多言語オントロジーに基づく
レシピ栄養分析に関する研究

鄭　夢龍

【修士論文概要書】

現在、世界中に不適当な飲食習慣や不規則な生活習慣などによって、糖尿病などの疾患や肥満のような健康問題が社会的問題となってきている。この問題を改善するために、より健康的な飲食生活を支援するサービスを考えている。本論文では、以下の五章に分けて、次世代のウェブ技術「セマンティック・ウェブ」やその中核の技術「オントロジー」に基づき、多言語レシピの抽出と栄養分析及び健康的なレシピ栄養評価・提案サービスの研究と解決方法を論じる。

１．概要

現在のインターネット技術の不足を指摘する、又は、健康管理の現状とレシピサイトの利用の際のレシピ栄養分析の難しさを説明する。更に、本研究の対象として、多言語オントロジーに基づくレシピの栄養分析と栄養評価・提案サービスを行う。

２．多言語データ検出オントロジーとは

オントロジーとは「人間から対象世界への根本的な理解」を介して「それらを体系的に書き記したもの」であると述べられている。分野ごとにこの定義に加えて若干異なる定義も説明されている。人工知能の立場では、「対象（世界）において興味を持つ概念とそれらの間の関係を明示化したもの」としている。

オントロジーの中のドメインオントロジーは、名前が示す通り特定の領域における対象の概念と概念間の関係を規定している。主に IS-A 関係と PART-OF 関係を明示する。例えば、自動車検索領域における中古車のオントロジーでは、「中古車」のメイン概念として、この下の「メーカー」、「車名」、「年式」、「価格」、「走行距離」などの概念が関係している。

現在のウェブ情報は単なる単語の集まりであり、コンピュータで自動的に処理されることまで考えられていなかった。オントロジーを導入すると、文書が大きな意味を持ったデータになって、計算機で意味に即して処理を行うことができる。それで、より知能的なサービスを提供することが可能になる。

各言語で記述されるウェブ情報を統合し、もっとも有効に利用するために、多言語オントロジーが提案された。多言語オントロジーは言語に関わらないオントロジー（A）と周りのN個ローカライゼーション（L_1 …L_n）から構成される。それぞれのローカライゼーションに各言語のオントロジーとマッピングセットを含める。マッピングセットによって、ローカライゼーションと言語に関わらないオントロジー（A）の相互マッピングができる。さらに、言語に関わらないオントロジー（A）を経由し、各ローカライゼーション間の相互マッピングが可能となる。マッピングの具体的な方法は三つある。単語を翻訳する時、「辞書マッチング」を利用する。通貨に対して、「通貨換算」を行う。異なる単位に対して、「単位換算」を行う。

３．多分野オントロジーとヒューリステイックスによる判断支援

人々は食生活をする時、好み、健康状況、文化、宗教などのことを考えた上で、レシピを決める。現在、多くの人がレシピサイトを利用し、レシピを検索している。しかし、今のレシピサイトはただレシピの情報を載せているにすぎず、レシピの検索やレシピの栄養分析、更にレシピの評価・提案に関するサービスは提供されていない。この問題を改善するために、本研究では、中国語、日本語、英語のレシピを対象にして、データ抽出オントロジーを用いて、レシピの要素を抽出し、栄養を計算する。これに基づき、レシピの評価・提案サービスを提案するために必要な一連の課題を研究する。

①**レシピ栄養オントロジーの定義**　栄養計算のため、材料名とその含量が必要である。そこで、オントジローの中に、「レシピ」、「材料特徴」、「材料名称」、「数量」、「計量単位」、「計量標準」、「栄養素」、「元素名」、「元素量」、「元素単位」という概念を定義した。多言語に対応するため、同じ形で中国語、日本語、英語のオントロジーを定義した。

②**レシピデータの抽出方法の定義**　レシピページの中から材料の情報だけを抽出する必要がある。しかし、ページ中に様々な内容があって、抽出に悪い影響を与える。そこで、前処理によって余分の内容を除く。前処理はサイト毎に定義され、DOMの解析でレシピの材料を記述する部分をマッピングする。処理済の情報からオントロジーで要素を抽出する。オントロジーの抽出方法は主に二つある。材料特徴や材料名称の辞書を用いて、レシピページから材料特徴や材料名称を抽出するという辞書マッチング方法と正規表現式のマッチングを使って、数量と計量単位を抽出するという正規表現マッチング方法である。レシピページにある数量と計量単位が様々な形式で表現されるので、複数の正規表現式を定義する。中国語、日本語、英語のオントロジーごとに辞書と正規表現式を定義する。

③**ローカライゼーション間のマッピング**　材料特徴、材料名称と数量、計量単位に分けてローカライゼーション間のマッピングを行う。材料特徴、材料名称をマッピングする場合、材料特徴、材料名称のRDF（Resource Description Framework）表現を使う。RDF表現は材料特徴、材料名称の各言語での表現を繋ぐ。RDF表現を使って、材料特徴、材料名称のマッピングを行う。数量、計量単位をマッピングする場合、単位RDF

表現と単位換算テーブルを使う。単位 RDF 表現は単位の各言語での表現と単位がどのローカルに属するかを示す。単位換算テーブルは標準単位とローカル単位の換算レートを示す。数量、計量単位をマッピングする時、先ず、マッピング先言語を用いて、マッピング先の単位をマッピングする。次に、マッピング元の単位とマッピング先の単位を使って、換算レートを取得する。最後に、換算レートを使って、マッピング先の単位に対応する数量を計算する。マッピング先の単位が複数ある場合に対応するため、本研究では単位選択ルールを定義する。このルールに従って、最も適当な単位を選択する。

④**レシピ栄養の計算**　健康的なレシピ評価・提案をするために、レシピの栄養データが必要である。本研究では、USDA（米国農務省）の栄養データに基づき、抽出されたレシピデータから栄養の計算を行う。レシピページにある材料名称が様々な形式で記述され、該当する栄養データを見つけられない可能性が高い。この問題を改善するために、抽出された材料特徴と材料名称の組合せを用いて、長い組合せから短い組合せへの順で、該当する栄養データをマッチングして、命中率を高めることを図る。又は、栄養データが 100g、1g あたりの材料の栄養含量を示しているが、レシピページ中に常に大匙、個など曖昧な単位がある。この問題を解決するため、材料毎の単位換算レートテーブルを定義する。このテーブルを用いて曖昧な単位を標準単位に換算し、材料の栄養を計算する。最後、各材料の栄養を合計し、レシピの栄養を計算する。

⑤**レシピ栄養評価・提案サービス**　レシピ栄養評価・提案をするために、健康状況や宗教より栄養摂取制限を定義する必要がある。本研究では、健康状況テーブル、栄養制限テーブル、禁止食材テーブルを定義し、栄養制限を表す。又は、レシピ栄養評価・提案ロジックが以下の順番で処理を行う。ユーザのクエリを解析し、ユーザの状況やレシピに関する検索条件を抽出する。次に、クエリ条件を作成し、オントロジー層に渡す。オントロジー層でレシピデータを取得し、目標言語に変換し、ロジック層に返す。次に、レシピ栄養評価・提案ロジックでユーザの状況に基づき、栄養摂取制限を推論する。最後に、栄養摂取制限を基準として、レシピの栄養を判断し、判断結果とレシピのデータをユーザに返す。

４．評価手順とその結果

レシピ栄養オントロジーの抽出精度が本研究にとって重要である。現段階で、日本語、中国語、英語のレシピサイトからレシピページを選択し、合計 202 個を対象にして、抽出の検証実験を行った。結果として、日本語レシピの抽出精度が 83.6%、中国語レシピの抽出精度が 85.6%、英語レシピの抽出精度が 87.1%、全体的な精度が 85.3%になっている。レシピページの中に複数の材料があったり、複数の単位があったり、表現が複雑だったせいで、約 15%のレシピが対応できなかった。

５．結論

現段階では、レシピデータの抽出検証はまだ不足である。将来、もっと多くのレシピページを対象にして抽出検証を行い、レシピ栄養オントロジーにおける用語の追加と抽出方法の改善によって、抽出精度を高めることができると考えている。本研究で論じたレシピ栄養評価・提案支援サービスを運用すると、人々の健康的な飲食生活に役立てることができるようになる。その結果、健康の促進や疾患の減少や政府医療支出の節約が期待できると思う。